Improving the reliability of, and confidence in, DFT functional benchmarking through active learning

Javier E. Alfonso-Ramos,[†] Carlo Adamo,[†] Éric Brémond,[*,‡] and Thijs Stuyver[*,†]

[†]*Ecole Nationale Supérieure de Chimie de Paris, Université PSL, CNRS, i-CLeHS, 75 005 Paris, France*

[‡]*Université Paris Cité, CNRS, ITODYS, 75 013 Paris, France*

E-mail: eric.bremond@u-paris.fr; thijs.stuyver@chimieparistech.psl.eu

Abstract

Validating the performance of exchange-correlation functionals is vital to ensure the reliability of DFT calculations. Typically, these validations involve benchmarking datasets. Currently, such datasets are typically assembled in an unprincipled manner, suffering from uncontrolled chemical bias, and limiting the transferability of benchmarking results to broader chemical space. In this work, a data-efficient solution, based on active learning, is explored to address this issue. Focusing – as a proof of principle – on pericyclic reactions, we start from the BH9 benchmarking dataset, and design a chemical space around this initial dataset by combinatorially combining reaction templates and substituents. Next, a surrogate model is trained to predict the standard deviation of the activation energies computed across a selection of 20 distinct DFT functionals. With this model, the designed chemical space is explored, enabling the identification of challenging regions, for which representative reactions are subsequently acquired. Remarkably, it turns out that the function mapping molecular structure to DFT functional divergence is readily learnable; convergence is reached upon the acquisition of less than 100 reactions. With our final model, a more challenging – and arguably more representative – pericyclic benchmarking dataset is curated, and we demonstrate that the functional performance has changed significantly compared to the original BH9 subset.

Introduction

Over the course of the past couple of decades, Density Functional Theory (DFT)[1,2] has grown into the workhorse of quantum chemistry and materials science, enabling a broad spectrum of applications.[3,4] An excellent accuracy/cost trade-off can in principle be achieved with the help of DFT, under the condition that a suitable exchange-correlation functional approximation (F_{xc}) is selected.[4,5] F_{xc} are usually ranked using Jacob's ladder,[6] where the most complex approximations can be found at the top. As a rule of thumb, functionals higher up the ladder tend to be more accurate, though this is most certainly not a universal rule that holds across application domains. As such, it is common practice in chemistry to determine the most suitable F_{xc} for a given type of application based on a benchmarking against a limited set of either experimental or high-level wave-function computed, data points.

One of the earliest examples of a benchmarking database was curated by Pople and co-workers, during their efforts to develop new quantum chemical methods for accurate energy calculations. Their database, dubbed "G1", consisted of a mere 31 atomization energies of atoms, ions, and organic molecules. This database later evolved into the G2/97, G3/99, and G3/05 databases;[7–10] the second incarnation of which was subsequently used to fit the empirical parameters of the popular B3LYP functional.[11] Since then, many more advanced and extensive – compiled – benchmarking datasets have been conceived, to allow the development and assessment of new methods across a large variety of chemical problems, e.g., Database 2015B by Thrular and co-workers,,[12] MGCDB84 database by Head-Gordon and co-workers,[4] and the GMTKN55 database by Grimme and co-workers.[13]

Despite the ever-expanding scope of modern benchmarking databases, it is not always clear to which extent the data points present in them are truly representative of the chemical spaces they are intended to describe. As previously noted by others,[14,15] the make-up of these datasets is often strongly biased, e.g., towards highly stable, easily accessible compounds for experimental datasets, and towards small, easily computable molecular systems for computational ones. Consequently, the transferability of exchange-correlation functional accuracy across chemical space is not inherently guaranteed, so a benchmarking study may lead to incorrect conclusions if no attention is paid to this issue. For example, it is not necessarily true that the functional approximation yielding the most accurate results on a specific small benchmarking set will also perform the best for the broader surrounding chemical space.[16] Even worse, while a benchmarking study on a small set of chemical systems for a given type of application may give the impression that a given set of functional approximations is sufficiently accurate – indicating that it should be possible to extract clear chemical trends from data computed at the corresponding levels of theory – this assumption may not hold if (some of) these same functionals struggle with specific regions of the chemical space under consideration that were not part of the original benchmarking set.

It is important to underscore that other researchers have previously thought about some of the issues outlined above, and several potential (partial) solutions have been proposed over the years.[14,17–20] One strategy that has been explored is the use of a recommender system, i.e., a machine-learning (ML) model that will predict which F_{xc} ought to perform best for the specific (molecular) system under consideration. Typically, a metric, indicative of the accuracy of the F_{xc} is designed, and then a multi-task regression model is trained.[17,18,21] Upon inference, the predicted scores of the individual functionals are ranked, and the model will recommend the best one. While arguably the most robust approach to deal with variations in functional performance across chemical space, it is important to appreciate the extreme computational cost of generating reliable training datasets for such recommender models. For example, DELFI, a recently developed recommender system for functional selection for excited state calculations of small organic molecules, required the construction of an initial dataset of 828 282 single-point TD-DFT calculations and over 21 000 reference values.[21]

An unrelated, and much less expensive, strategy to reduce the odds of suboptimal DFT functional selection consists of 'mindless benchmarking'.[14] This approach, pioneered by Grimme and co-workers, involves random generation of artificial molecular geometries, after which a representative set is sampled with a particular focus on diversity. In this manner, the introduction of biases during the benchmarking dataset construction is limited, increasing the likelihood that the resulting set is at least somewhat representative of the broader chemical space that it aims to describe. Despite the promise of this approach for benchmarking some specific chemistry-related tasks more reliably, the strategy is not ideally suited for others. For example, when benchmarking reaction kinetics, one typically aims to identify a good functional approximation for specific reaction classes, across a well-defined scope of the accessible chemical space, so that a purely randomized approach is hard to implement.

In this work, a principled, data-driven strategy to assess the transferability of benchmarking results and to construct more challenging and/or representative datasets will be presented, based on active learning. With a particular focus on the simulation of pericyclic reactions, we train here a machine learning model on a small pre-existing benchmarking dataset, to identify regions of chemical space for which activation energies computed with different functional approximations diverge the most. Next, a Bayesian optimization (BO) algorithm[22–25] is applied to identify the reactions in the chemical space surrounding the initial dataset exhibiting the biggest variations in DFT computed activation energies. The selected reactions then ought to be representative of more challenging patches of chemical space than those present in the training set.

Iteratively acquiring these reactions and updating the model accordingly, we demonstrate that its generalizability across the defined chemical space improves rapidly, i.e., the discrepancy between predicted and computed variations in activation energy values of the acquired points gradually declines, and convergence is reached after only a handful iterations, corresponding to the acquisition of fewer than 100 new data points. With the final, validated model at our disposal, new benchmarking datasets with particularly challenging reactions can be curated, and an informed estimate of the maximal errors across chemical space can be inferred.

Overall, we believe that the presented approach provides a computationally inexpensive, and highly data-efficient, strategy to improve the reliability of – and confidence in – benchmarking efforts. Furthermore, our strategy is easily extensible to other regions of the chemical space, so that the presented work could provide a blueprint for further advances in simulation method selection for future chemical reactivity studies, as well as in the development of new, robust DFT functionals.

Methodology

Extracting and curating the initial data from the BH9 dataset

BH9 is an extensive and diverse benchmark dataset for reaction and activation energies, composed of 449 chemical reactions belonging to nine types common in organic chemistry and biochemistry.[26] The molecular species in BH9 comprise main-group elements (H, C, N, O, F, P, S, and Cl), plus B and Si. The pericyclic subset, on which we will focus throughout this study, consists of 140 Diels-Alder (DA), [3+2] cycloaddition (DC), electrocyclic, [3,3] rearrangement (RR), [6+4], [4+6], [8+2] and [2+2+2] cycloaddition reactions. We focus on this specific reaction class because of its chemical and biological importance, as well as the good understanding and relative robustness of the corresponding mechanisms.[27–29]

Target selection

Within this work, we aim to identify increasingly challenging reactions, i.e., reactions for which the activation energy exhibits the highest variability across several functionals. Two common quantities can in principle guide us to this end, the range and the standard deviation

(σ) of the computed activation energies. We considered the range to be less useful here, as it does not convey any information regarding how dispersed the values truly are, i.e., a high range can be obtained because a single functional is an outlier for a given reaction data point, while this datapoint may still result in a narrow distribution of (activation) energy values overall. σ on the other hand is a direct measure of how dispersed the values are around the mean. For this reason, we selected it as the target quantity throughout this study.

Generation of the chemical space

The first step towards the design of a broader chemical space around the benchmarking dataset consisted of converting the geometries in the pericyclic subset of BH9 into SMILES, using the functionalities of RDKit.[30] Subsequently, reactive cores, i.e., collections of atom pairs undergoing a change in bonding throughout the reaction, were selected. Only those cores with a repeated occurrence, and for which a high σ in the DFT computed activation energies across functionals is obtained (a cutoff of 4.80 kcal/mol was set), were selected. Both the mentioned criteria were considered essential. As an illustration, the [8+2] addition exhibits the highest standard deviation in the activation energy values across the BH9 subset (7.52 kcal/mol), but this type of reaction occurs only once (in two different regioisomers), and hence this reaction subclass was rejected. On the other hand, 53 data points of the Diels-Alder [4+2] addition are present in BH9, with the reaction between naphthoquinone and a functionalized 1,3 butadiene being the most divergent with a standard deviation in the activation energy over 6.50 kcal/mol (First core of the Diels-Alder box in Fig. 1).

In total, 10 reactive cores were selected as templates in this manner, respectively divided into five DA, three DC, and two RR ones (the corresponding reaction SMILES can be found in Section S1 of the Supporting Information). An initial set of 9 substituents to decorate the selected reactive cores were identified from the same subset of the BH9-extracted pericyclic reactions. Additionally, 5 extra cores were included to enable the generation of fused rings. In these cores, both reactants are connected with a linker to create new fused rings with a size ranging from 3 to 8 members (Figure 1). With these structural elements, a total of 9045 reactions could be generated, constituting our chemical space.

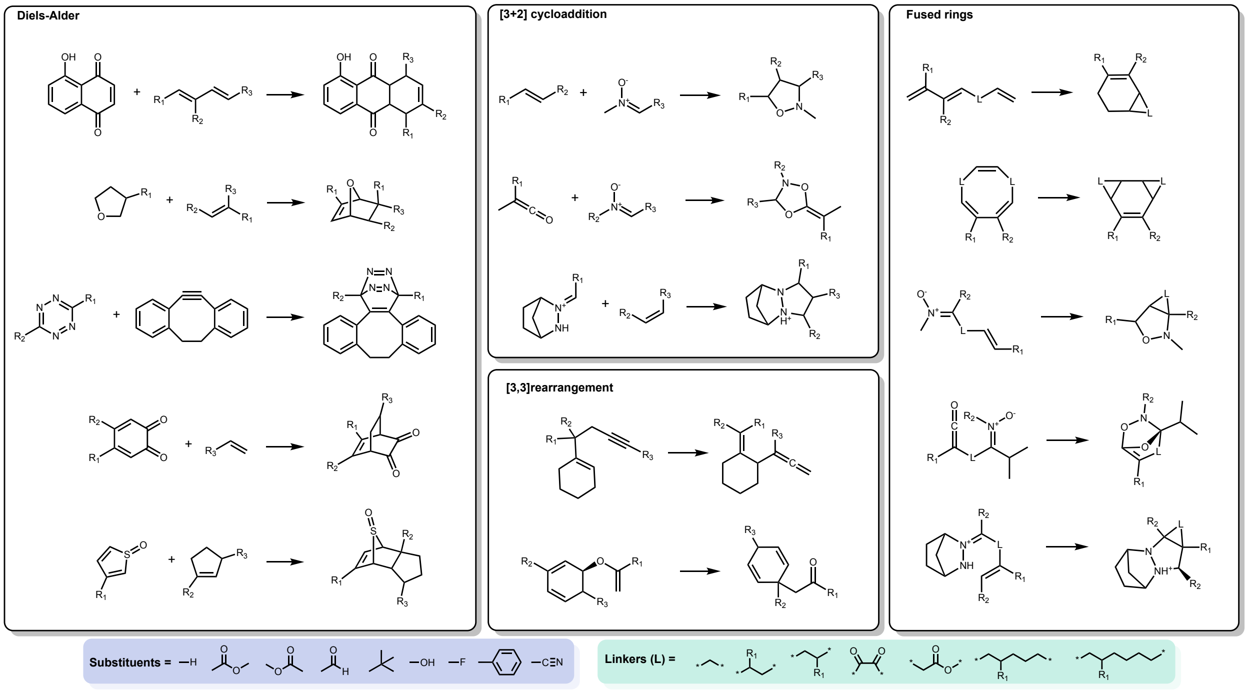


Figure 1: Schematic overview of the designed chemical space. The substituents are in the purple box, and the linkers are in the green box.

Fingerprints

Two types of fingerprints were investigated as potential reaction representations, the differential reaction fingerprints (DRFP),[31] and the differential Morgan fingerprints.[32] Both belong to the family of circular fingerprints, where the chemical information of the surrounding atoms up to a radius r is encoded as a machine-readable representation. The DRFP algorithm creates a binary fingerprint based on the symmetric difference of two sets, containing the circular molecular n-grams generated from the molecules listed left and right from the reaction arrow, respectively. The Morgan fingerprints of the reaction SMILES were computed by subtracting the molecular fingerprints of the reactants from those of the products.

A limitation of fingerprint-derived representations is that they are inherently local in nature, i.e., they do not capture long-range changes in the molecular structure. Since it became clear from a preliminary analysis of the BH9 subset that such phenomena, e.g., the formation or disintegration of a macrocycle throughout the reaction, can have an outsized effect on the functional approximation divergence (*vide infra*), we thus decided to concatenate a final 6-bit vector, encoding the information about the size of new rings formed/broken. Overall, the fingerprints were generated using radius $r = [1, 2, 3]$ and dimensionality $dim = [256, 512, 1024, 2048]$.

Computational details

A fully automated reaction profile computation workflow based on autodE[33] and Gaussian16[34] was set up to acquire new reaction data points. As the level of theory for the DFT calculations, CAM-B3LYP[35] in combination with the 6-311++G** basis set in the gas phase was used for the final optimization of all stationary points. The maximum number of conformers generated for single molecules and transition states was set to 1000, using an RMSD cut-off of 0.1 Å to exclude identical conformers. Conformers were ranked based on a loose optimization at CAM-B3LYP/6-31G*, and the lowest energy one was selected for refinement with the larger basis set. The D3 dispersion correction with Becke-Johnson damping was used in all cases.[36] An IRC confirmation at the same level of theory was consistently performed on the final TS. The final functional and basis set were selected with the aim of approaching the level of theory of Prasad *et al.*[26] as closely as possible.

A set of 20 exchange-correlation functionals, listed in Table 1 has been considered in this study. Single-point energy calculations were carried out using the def2-QZVPP[37,38] basis set on the optimized molecular structures of reactants, products, and transition states. The RIJCOSX approximation[39] was always considered for Coulomb and exchange integrals together with the auxiliary basis set automatically[40] generated by ORCA. The reference energies were computed at DLPNO-CCSD(T) level using a TightPNO selection threshold. Both types of calculations were performed using the 5.0.4 release of ORCA,[41,42] and selecting the DefGrid3 integration grid by default, as well as a very tight SCF convergence criterion.

More details about the autodE methodology and reference calculations can be found in Section S2 of the ESI.

Table 1: List of the exchange-correlation functionals considered in this work, ranked according to the casted percentage of the exact-like exchange (EXX) and second-order perturbation theory (PT2) correlation contributions.

Functional	% EXX	% PT2	Ref.
GGA and meta-GGA			
BLYP	0	0	[43,44]
M06-L	0	0	[45]
B97M-V	0	0	[46]
Global hybrids			
B3LYP	20	0	[11,43]
PBE0	25	0	[47]
M06	27	0	[48]
BHandHLYP	50	0	[43]
M06-2X	54	0	[48]
Range separated hybrids			
ωB97M-V	15/100	0	[49]
ωB97X	15.8/100	0	[50]
ωB97X-V	16.7/100	0	[51]
CAM-B3LYP	19/65	0	[35]
Double hybrids			
PBE0-DH	50	12.5	[52]
B2-PLYP	53	27	[53]
PBE-QIDH	69.3	33.3	[54]
B2K-PLYP	72	42	[55]
Range separated double hybrids			
RSX-0DH	50/100	12.5	[56]
ωB2-PLYP	53/100	27	[57]
ωB2-PLYP18	53/100	27	[57]
RSX-QIDH	69.3/100	33.3	[56,58]

Bayesian optimization

Bayesian optimization (BO) is an adaptive procedure to efficiently obtain a global maximum (or minimum) of a function f, of which the analytic form or its derivatives are not known, and whose evaluation tends to be expensive with respect to time, resources, and/or budget.

In mathematical terms, the problem can be summarised as follows:

$$x^* = argmax_{x \in X}\ f(x) \tag{1}$$

where x^* is the point in the input representation space that produces the maximum of the function.

BO is a direct application of Bayes Theorem,[59] and consists of two main components: the construction of a surrogate model to approximate the black-box function f to be optimized, and an acquisition function for deciding the next samples to evaluate. Pseudo-code for the BO algorithm used can be found in the ESI; for more details about the foundations and/or applications, we refer to references.[60,61]

Surrogate model

A surrogate model is a trained ML model used to predict the objective function of the reaction in the generated chemical space. Three types of surrogate models were explored, namely K-Nearest Neighbor, Random Forest, and XGBoost. The first two were implemented in Python using the packages Scikit-learn,[62] and the latter has been implemented with the help of a dedicated package, XGBoost.[63] More details regarding the architectures can be found in Section S5.1 of the ESI.

Acquisition function

The acquisition function is used to select the most promising evaluations to perform next, i.e., which reaction profile to acquire with the help of our automated reaction profile computation workflow. Acquisition functions are typically derived by considering the mean, $\mu(x)$, and standard deviation, $\sigma(x)$, of the surrogate model predictions, either by considering an ensemble of ML models trained on various distinct data splits, or by analyzing the distribution of the individual estimator predictions (e.g., when random forests are selected). In this work, the upper confidence bound (UCB) has been used specifically:

$$UCB(x) \;=\; \mu(x) \; + \beta\sigma(x) \tag{2}$$

where β is an explicit hyperparameter to balance the ratio between exploitation and exploration, higher beta values will result in the prioritization of zones for which the model is uncertain, i.e., exploration is favored, while lower values will result in the prioritization of sampling data points with high predicted performance, i.e., exploitation is favored.

Results and discussion

Exploratory analysis

First, we started with an exploratory data analysis of the pericyclic subset of BH9 and the connection between specific structural elements and the observed variation in the activation energy values computed across the set of exchange-correlation functionals probed. Figure 2a illustrates that DA is the most common reaction type in the dataset, followed by electrocyclic reactions, RR, and DC. At first glance, it seemed that reactions involving highly conjugated systems exhibited a more significant activation energy variability across several functionals, while intramolecular reactions appeared to exhibit a less pronounced variability (Fig. 2c). Additionally, other factors, such as a significant dependence on the direct substitution of the reactive atoms, and steric effects can also be discerned.

From Fig. 2a, it should be obvious that, for the Diels-Alder reactions present in the BH9 dataset in particular, the performance of the different functionals is highly heterogeneous. For some reactions, σ is negligible, i.e., all functionals agree rather well in terms of the computed magnitude of the activation energy, while for other reactions, the divergence across the different functionals is much more pronounced.

This observation brings us back to the core of the problem we aimed to sketch in the introduction: selecting benchmarking reactions in an unprincipled manner can easily result in profoundly unrepresentative and misleading conclusions about functional performance and robustness.

To further drive home this point, we split the Diels-Alder reactions present in BH9 into two hypothetical subsets, which we call here benchmarking subsets 1 and 2 respectively. Benchmarking subset 1 consists of all the reactions with σ below 3.5 kcal/mol; subset 2 consists of all reactions with a higher σ. Comparing the performance of the different functionals, with respect to the DLPNO-CCSD(T) computed reference, between the two subsets, we observe remarkable differences, both in qualitative and quantitative terms (Fig. 2b). While the 3 best-performing functionals, ωB2-PLYP, ωB97M-V, and M06-2X are only negligibly affected by evaluating a different subset, the ordering of the remaining functionals changes profoundly. For example, the functional ranking in fourth place in terms of performance for benchmarking subset 1, ωB97X-V, drops 3 places when evaluated on subset 2. Remarkably, while BLYP and B3LYP are seemingly quite robust when evaluated on subset 1, appearing halfway in the functional ranking and resulting in an acceptable MAE of 2-3 kcal/mol, for subset 2, their mean errors almost triple (to 8-9 kcal/mol), rendering these functionals the second and fourth least reliable functionals of all the ones tested, respectively. Adding empirical dispersion corrections partially remedies the remarkable failure of these functionals, but even then, significant performance losses are still observed (cf. Section S3 in the ESI).

The observation above is particularly concerning because the latter functionals are still commonly used in reactivity studies. Oftentimes, a handful of benchmarking data points, indicating that these functionals do not perform dramatically worse than more modern functionals, are used as justification for their selection, as the BLYP and B3LYP functionals tend to also be among the fastest to evaluate.[64,65] This preliminary analysis however already demonstrates that caution is needed in this regard, as trends emerging in local patches of chemical space may not always hold beyond them.

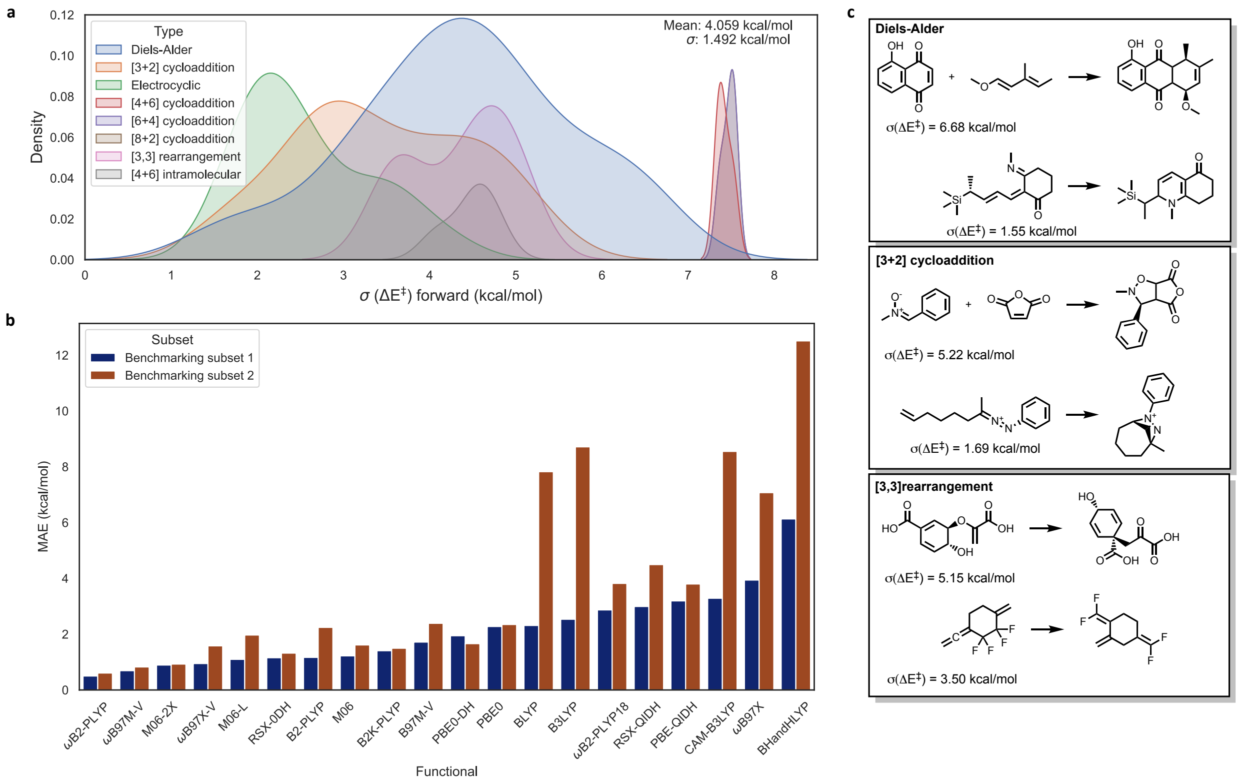


Figure 2: (a) Kernel density estimate plots show the distribution of the target value for every reaction class. (b) Mean Absolute Errors (MAE, kcal/mol) with respect to the DLPNO-CCSD(T) reference, for the activation energy for the two benchmarking subsets, computed using the 20 DFT approaches considered in the present paper. (c) Examples of reactions with the highest and lowest standard deviation for DA, DC, and RR.

Reaction representation and surrogate model

As described in Section 3, three surrogate model architectures in combination with two different fingerprint representations were explored. We evaluated the performance of each surrogate model/fingerprint combination using 6-fold nested cross-validation on the 140 cycloaddition reactions of the original BH9 dataset. Morgan fingerprints clearly outperform DRFP, a lower radius (r=1 and r=2) results in lower errors than higher ones (r=3), and increasing the number of bits, decreases the error in our calculations. Table 2 summarizes the performance of the three best models trained. Results of all tested models and more details about hyperparameter optimization can be found in Section S5.2 of the Supporting Information. Based on these results, a random forest together with Morgan fingerprint with r=2 and bits=2048 were selected as the most suitable surrogate model and molecular representation combination.

Table 2: A summary of the performance of the three best model architectures tested on the original BH9 pericyclic dataset.

model	fingerprint	r	bits	RMSE (kcal/mol)	MAE (kcal/mol)	R^2
RF	Morgan	2	2048	0.647	0.500	0.795
RF	DRFP	1	1024	0.669	0.502	0.782
RF	Morgan	1	2048	0.652	0.508	0.793

Bayesian optimization campaign

Having chosen the previously mentioned reaction representation/surrogate model and defined the chemical space, the BO campaign was initiated. For the initial rounds, a β value of 1.8 was set to favor the exploration of the unknown regions, and a maximum of 15 points were acquired per round. Besides the acquisition function, we also applied a filter based on structural diversity, i.e., we required a minimal cosine distance between the fingerprints of new reactions being selected for acquisition. By applying this diversity filter, we ensured that the various reactive cores were each sampled as part of the campaign.[66] The motivation for this approach stems from the observation that some cores intrinsically exhibit a higher σ than others; e.g. the highest value for the Diels-Alder reactions amounts to 6.68 kcal/mol, while for the [3,3] rearrangements, the highest value in the original BH9 subset amounts to a mere 5.15 kcal/mol. Without diversity-based filtering, we would have sampled almost exclusively reactions for the reaction cores resulting in the most divergent activation energy values in the initial dataset, and only minimal exploration would have been performed for the other cores at best. The settings of each acquisition round can be found in Section S2.1 of the ESI (additionally, a bash script reproducing each round is available in our GitHub repository).[67]

The performance of our model after each acquisition round was monitored with the help of nested 6-fold cross-validation. We observed that the (intrapolation) MAE consistently oscillated around the initial value of 0.52 kcal/mol (R^2=0.78), and did not improve throughout the campaign (*cf.* Figure 3a). This however does not mean our model was not improving overall/becoming more robust: the accuracy of the predictions across the wider constructed chemical space rose rapidly, indicating improvements in the generalizability power across the chemical space of the model. From 1.15 kcal/mol in round one, the MAE on σ of the activation energy values gradually and monotonously decreased to 0.60 kcal/mol by round seven, i.e., the mean error was cut by over half its initial value. Starting from round five, a smooth leveling-off in the error reduction could be observed, and the prediction error in the final round approached the error obtained during (intrapolative) cross-validation. Both of these observations suggest that convergence had essentially been reached by this point.

It should be noted here that the fact that a model, trained on a combination of the original BH9 subset and the acquired reactions, generalizes much better across the designed chemical space, suggests that this resulting, expanded dataset is significantly more representative of the selected chemical space than a (subset of) the original benchmarking dataset.

Figure 3c presents the distribution of the σ values for the newly acquired points in each round, with the baseline, i.e., the highest σ values recorded for each reaction class in the original pericylic subset of BH9, indicated as dashed lines.

Already in the second round, the baseline values are exceeded for each subclass, with the [3+2] cycloadditions resulting in the highest jump, with an increase of σ by 2.13 kcal/mol. Note that this plot also reflects the previously mentioned gradual improvement of the model throughout the campaign: in the first three rounds, the computed divergence in activation energy values across functionals for the acquired points are scattered across a wide range (σ values range from 2-7 kcal/mol), whereas starting in round four, and particularly in round six/seven, the computed σ values for new acquisitions become much more concentrated towards the higher end of the spectrum of σ values, and approach or exceed the top values

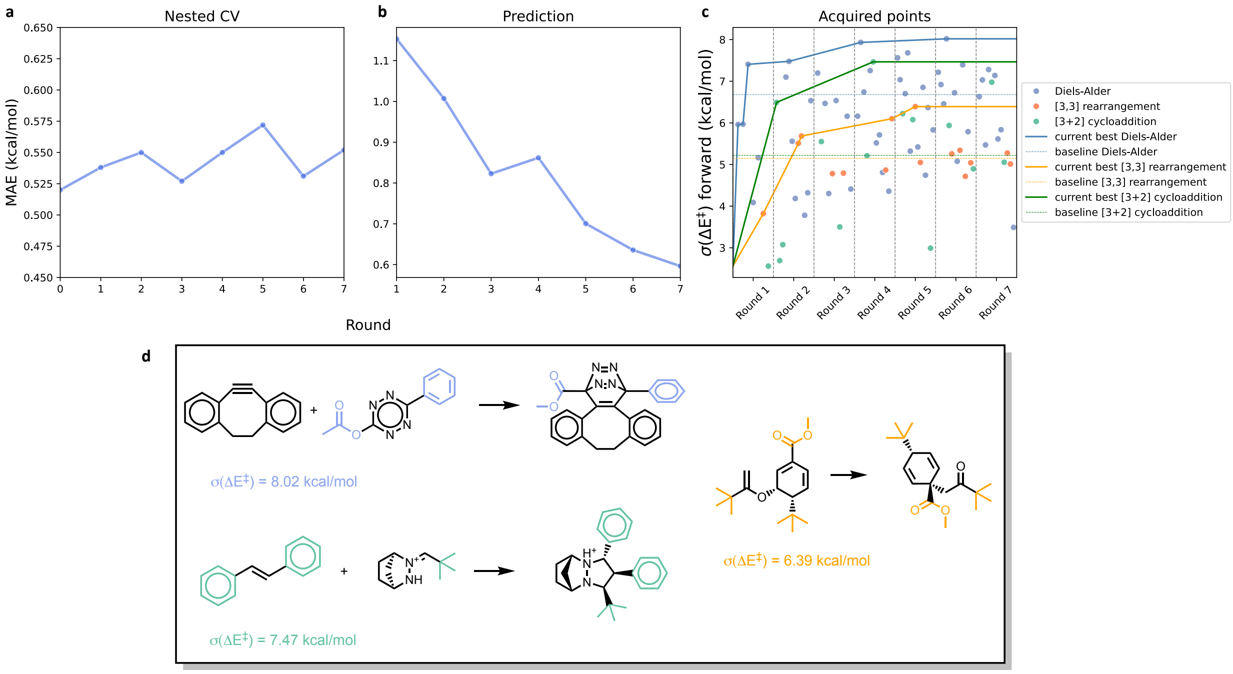


Figure 3: Mean Absolute Errors (MAE, kcal/mol) for the target value during (a) nested cross-validation of the training data and (b) acquired data points. (c) Plot of the σ activation energies of each acquired data point during each round; bold curve represents the best values and horizontal dashed lines the best baseline values. (d) Current best reactions after 7 rounds of BO campaign; substitutions groups are highlighted. Values are shown in kcal/mol.

present in the original dataset (some dispersion inevitably remains due to the diversity-based filtering, *vide supra*).

As already indicated, after 7 rounds, the exploration stage of the BO campaign was stopped, as the σ of the newly acquired data points had barely improved with respect to the previous rounds, and the MAE of the predictions appeared to approach convergence. Some examples of the best reactions emerging from the exploration stage are shown in Figure 3d, with the selected substituents highlighted. Substituents that extend the delocalization, such as phenyl or ester groups, as well as voluminous groups introducing steric hindrance, turn out to be key for increasing the σ of the activation energy values.

Upon conclusion of the exploration stage, a final exploitation round, consisting of the acquisition of 30 points (with a β value of 1.0), was initiated. From this batch, the reaction profiles were successfully acquired for 26 reactions. The MAE of the predictions on this round was 0.61 kcal/mol, in line with the error obtained in the last exploration round, confirming that model convergence had indeed been reached. From the final pool of 108 acquired reactions, 70 were selected to form our new benchmarking dataset. A table containing the list of reactions for this new dataset organized by type, and the reference reaction energy and barrier heights can be found in Section S2.4 of the ESI.

DLPNO-CCSD(T) reference values were computed for the acquired reactions (see Section S2.4 of the ESI for the methodology used), and the deviation for every functional was computed. The resulting accuracies are presented graphically in Figure 4, together with the accuracies for the original BH9 cycloaddition dataset (values are presented in Table S4 of ESI).

Overall, the individual errors, as well as the relative ordering of the functionals have changed dramatically. Three functionals that appeared in the middle of the pack when evaluated on the original BH9 subset, perform best on our newly acquired data points: ωB2-PLYP18, PBE-QIDH, and RSX-QIDH. Remarkably, these three functionals buck the trend of all the other functionals, by actually performing better on the new data compared to the original data.

For most of the other functionals, the deterioration in performance is far from uniform, and at times spectacular. For example, M06, PBE0, and B2-PLYP, which appeared fairly robust when subdividing the Diels-Alder reactions of BH9 into two benchmarking subsets (cf. Figure 2b), exhibit a mean error at least three times larger when evaluated on our active learning dataset instead of on BH9. In absolute terms, the deterioration of BLYP, CAM-B3LYP and BHandHLYP is even worse, but as indicated before, adding dispersion corrections remedies the situation to some extent (cf. Section S3 of the ESI). The ωB97M-V and M06-2X functionals on the other hand remain relatively robust, though the MAEs still deteriorate by approximately 2 kcal/mol. Because of this deterioration, the former functional, which performed best on the original BH9 subset, now only comes in fourth place when ranking the functionals based on performance.

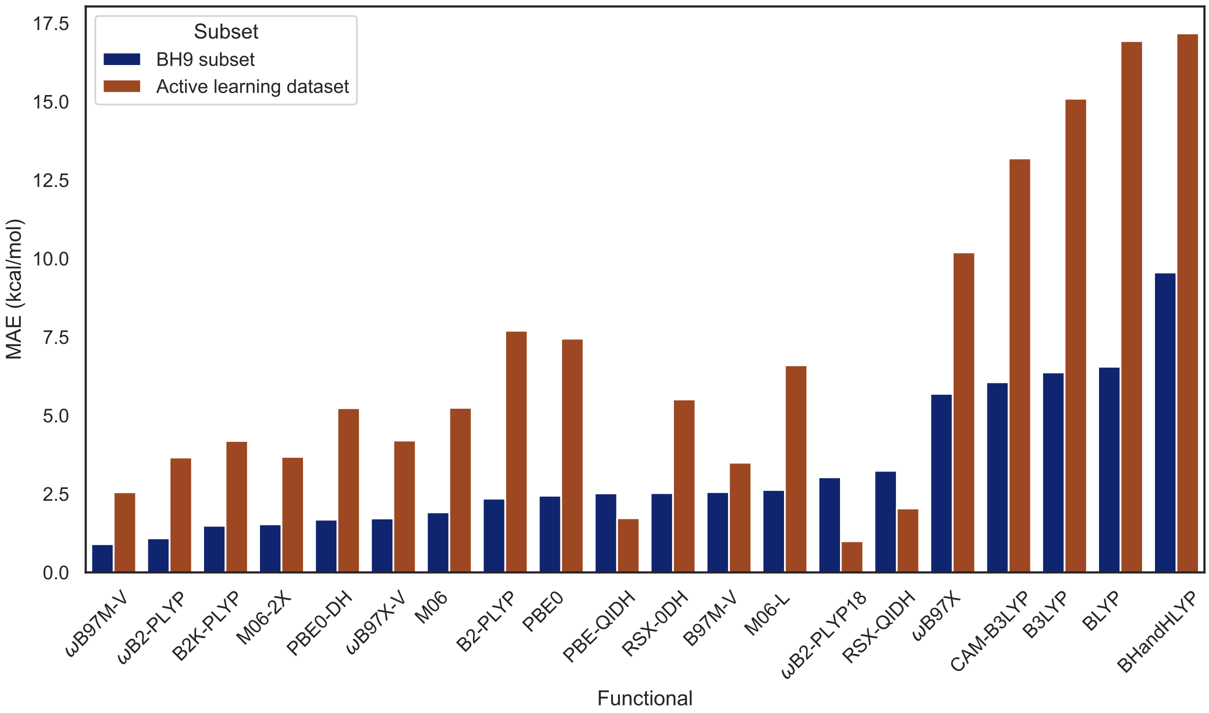


Figure 4: Mean Absolute Errors (MAE, kcal/mol) with respect to the DLPNO-CCSD(T) reference, for the activation energy for the original BH9 subset and our subset, computed using the 20 DFT approaches considered in the present paper.

Conclusions

Validation of exchange-correlation functionals is an essential task in (computational) chemistry, and the design of benchmark datasets is key to this end. In this work, a new approach, based on Bayesian optimization and active learning, for generating more diverse, unbiased, benchmarking datasets is presented. Starting from an initial model trained on 140 data points, our strategy enables the identification of challenging regions of chemical space within only a handful of iterations. Sampling new data points in these regions reveals that we have successfully pushed the mean errors for most functionals higher, and a completely different picture of the performance is obtained, with the ranking of some functionals being changed dramatically. It is important to note that the strategy followed in this paper is easily extensible to other relevant chemical properties (as well as functionals), and is extremely data-efficient.

While we consider this work in the first place as a proof of principle, instead of a comprehensive benchmarking study in its own right, we can nevertheless interpret the obtained results and provide some guidelines about which functionals to select when studying cycloaddition reactions. ωB2-PLYP18, and PBE-QIDH turn out to be extremely robust functionals, exhibiting MAEs on our set of acquired reactions lower than 2.0 kcal/mol. Alternatives (and cost-effective) options could be the hybrid ωB97M-V, and the meta-GGA B97M-V with an MAE of 2.6 and 3.5 kcal/mol, respectively.

The main conclusion of this study, however, is that current benchmarking datasets such as BH9 are not necessarily representative of their surrounding chemical spaces. This is an issue that has received limited attention up to this point but needs to be addressed to improve the reliability of – and the confidence in – DFT studies.

Data availability

The code used to generate/curate the dataset as well as the Bayesian optimization and active learning procedure can be found at https://github.com/chimie-paristech-CTM/ML_DFT_benchmarking. The (generated/curated) datasets can be downloaded at https://figshare.com/projects/ML_DFT_benchmarking/219457.

Author Contributions

JEAR: data curation, software, methodology, formal analysis, visualization, writing. CA: conceptualization, formal analysis, writing, methodology. EB: conceptualization, software, formal analysis, writing, methodology. TS: conceptualization, software, methodology, formal analysis, writing, supervision, funding acquisition.

Conflicts of interest

There are no conflicts to declare.

Acknowledgement

JEAR and TS acknowledge the French National Agency for Research (ANR) for a CPJ grant (ANR-22-CPJ1-0093-01). This work was granted access to the HPC resources of IDRIS under the allocation *2023-100732* and of TGCC under the allocation *AD010810359R3* granted by GENCI. E. B. acknowledges the ANR MoMoPlasm project (ANR-21-CE29-0003), and Labex SEAM (Science and Engineering for Advanced Materials and devices) ANR-10-LABX-096, ANR-18-IDEX-0001. C. A. acknowledges funding by the European Union (ERC AdG, project MaMa, no. 101097351). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Supporting Information Available

- SMILES templates, computational details, reference energies, a table containing the list of reactions for our active learning benchmarking dataset organized by type and performance of surrogate models.

References

(1) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864–B871.

(2) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133–A1138.

(3) Huang, B.; von Rudorff, G. F.; von Lilienfeld, O. A. The central role of density functional theory in the AI age. *Science* **2023**, *381*, 170–175.

(4) Mardirossian, N.; Head-Gordon, M. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Mol. Phys.* **2017**, *115*, 2315–2372.

(5) Goerigk, L.; Grimme, S. A thorough benchmark of density functional methods for general main group thermochemistry, kinetics, and noncovalent interactions. *Phys. Chem. Chem. Phys.* **2011**, *13*, 6670–6688.

(6) Perdew, J. P.; Ruzsinszky, A.; Tao, J.; Staroverov, V. N.; Scuseria, G. E.; Csonka, G. I. Prescription for the Design and Selection of Density Functional Approximations: More Constraint Satisfaction with Fewer Fits. *J. Chem. Phys.* **2005**, *123*, 062201.

(7) Pople, J. A.; Head-Gordon, M.; Fox, D. J.; Raghavachari, K.; Curtiss, L. A. Gaussian-1 theory: A general procedure for prediction of molecular energies. *J. Chem. Phys.* **1989**, *90*, 5622–5629.

(8) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. Assessment of Gaussian-2 and density functional theories for the computation of enthalpies of formation. *J. Chem. Phys.* **1997**, *106*, 1063–1079.

(9) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. Assessment of Gaussian-3 and density functional theories for a larger experimental test set. *J. Chem. Phys.* **2000**, *112*, 7374–7383.

(10) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Assessment of Gaussian-3 and density-

functional theories on the G3/05 test set of experimental energies. *J. Chem. Phys.* **2005**, *123*, 124107.

(11) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

(12) Haoyu, S. Y.; Zhang, W.; Verma, P.; He, X.; Truhlar, D. G. Nonseparable exchange–correlation functional for molecules, including homogeneous catalysis involving transition metals. *Phys. Chem. Chem. Phys.* **2015**, *17*, 12146–12160.

(13) Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Phys. Chem. Chem. Phys.* **2017**, *19*, 32184–32215.

(14) Korth, M.; Grimme, S. "Mindless" DFT Benchmarking. *J. Chem. Theory Comput.* **2009**, *5*, 993–1003.

(15) Chan, B.; Dawson, W.; Nakajima, T. Data Quality in the Fitting of Approximate Models: A Computational Chemistry Perspective. *J. Chem. Theory Comput.* **2024**, *20*, 10468–10476.

(16) Li, H.; Mansoori Kermani, M.; Ottochian, A.; Crescenzi, O.; Janesko, B. G.; Truhlar, D. G.; Scalmani, G.; Frisch, M. J.; Ciofini, I.; Adamo, C. Modeling Multi-Step Organic Reactions: Can Density Functional Theory Deliver Misleading Chemistry? *J. Am. Chem. Soc.* **2024**, *146*, 6721–6732.

(17) Liu, F.; Duan, C.; Kulik, H. J. Rapid Detection of Strong Correlation with Machine Learning for Transition-Metal Complex High-Throughput Screening. *J. Phys. Chem. Lett.* **2020**, *11*, 8067–8076.

(18) Meyer, R.; Arunachalam, N.; Kulik, H. J. A transferable recommender approach for selecting the best density functional approximations in chemical discovery. *Nat Comput Sci* **2023**, *3*, 38–47.

(19) Gould, T.; Chan, B.; Dale, S. G.; Vuckovic, S. Identifying and embedding transferability in data-driven representations of chemical space. *Chem. Sci.* **2024**, *15*, 11122–11133.

(20) Khan, D.; Price, A. J. A.; Ach, M. L.; von Lilienfeld, O. A. Adaptive hybrid density functionals. *arXiv preprint arXiv:2402.14793* **2024**,

(21) Avagliano, D.; Skreta, M.; Arellano-Rubach, S.; Aspuru-Guzik, A. DELFI: a computer oracle for recommending density functionals for excited states calculations. *Chem. Sci.* **2024**, *15*, 4489–4503.

(22) Mockus, J. *Bayesian Approach to Global Optimization: Theory and Applications*; Mathematics and its Applications; Springer Netherlands, 2012.

(23) Moosavi, S. M.; Jablonka, K. M.; Smit, B. The Role of Machine Learning in the Understanding and Design of Materials. *J. Am. Chem. Soc.* **2020**, *142*, 20273–20287.

(24) Häse, F.; Roch, L. M.; Kreisbeck, C.; Aspuru-Guzik, A. Phoenics: A Bayesian Optimizer for Chemistry. *ACS Cent. Sci.* **2018**, *4*, 1134–1145.

(25) Ranković, B.; Griffiths, R.-R.; Moss, H. B.; Schwaller, P. Bayesian optimisation for additive screening and yield improvements – beyond one-hot encoding. *Digit. Discov.* **2024**, *3*, 654–666.

(26) Prasad, V. K.; Pei, Z.; Edelmann, S.; Otero-de-la Roza, A.; DiLabio, G. A. BH9, a New Comprehensive Benchmark Data Set for Barrier Heights and Reaction Energies: Assessment of Density Functional Approximations and Basis Set Incompleteness Potentials. *J. Chem. Theory Comput.* **2022**, *18*, 151–166.

(27) Carruthers, W. *Cycloaddition Reactions in Organic Synthesis*; Organic chemistry series; Elsevier Science, 1990.

(28) Sankararaman, S. *Pericyclic Reactions - A Textbook: Reactions, Applications and Theory*; Wiley, 2005.

(29) Jamieson, C. S.; Ohashi, M.; Liu, F.; Tang, Y.; Houk, K. N. The expanding world of biosynthetic pericyclases: cooperation of experiment and theory for discovery. *Nat. Prod. Rep.* **2019**, *36*, 698–713.

(30) RDKit: Open-source cheminformatics. `http://www.rdkit.org`.

(31) Probst, D.; Schwaller, P.; Reymond, J.-L. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digit. Discov.* **2022**, *1*, 91–97.

(32) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(33) Young, T. A.; Silcock, J. J.; Sterling, A. J.; Duarte, F. autodE: automated calculation of reaction energy profiles—application to organic and organometallic reactions. *Angew. Chem., Int. Ed.* **2021**, *133*, 4312–4320.

(34) Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V.; Petersson, G.; Nakatsuji, H.; others Gaussian 16. 2016.

(35) Yanai, T.; Tew, D. P.; Handy, N. C. A new hybrid exchange–correlation functional using the Coulomb-attenuating method (CAM-B3LYP). *Chem. Phys. Lett.* **2004**, *393*, 51–57.

(36) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

(37) Weigend, F. Accurate Coulomb-fitting basis sets for H to Rn. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057–1065.

(38) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(39) Neese, F.; Wennmohs, F.; Hansen, A.; Becker, U. Efficient, approximate and parallel Hartree–Fock and hybrid DFT calculations. A 'chain-of-spheres' algorithm for the Hartree–Fock exchange. *Chem. Phys.* **2009**, *356*, 98–109.

(40) Stoychev, G. L.; Auer, A. A.; Neese, F. Automatic Generation of Auxiliary Basis Sets. *J. Chem. Theory Comput.* **2017**, *13*, 554–562.

(41) Neese, F. The ORCA program system. *WIRES Comput. Molec. Sci.* **2012**, *2*, 73–78.

(42) Neese, F. Software update: the ORCA program system, version 5.0. *WIRES Comput. Molec. Sci.* **2022**, *12*, e1606.

(43) Becke, A. D. A new mixing of Hartree–Fock and local density-functional theories. *J. Chem. Phys.* **1993**, *98*, 1372–1377.

(44) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785–789.

(45) Zhao, Y.; Truhlar, D. G. A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. *J. Chem. Phys.* **2006**, *125*, 194101.

(46) Mardirossian, N.; Head-Gordon, M. Mapping the genome of meta-generalized gradient approximation density functionals: The search for B97M-V. *J. Chem. Phys.* **2015**, *142*, 074111.

(47) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(48) Zhao, Y.; Truhlar, D. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: Two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

(49) Mardirossian, N.; Head-Gordon, M. ωB97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation. *J. Chem. Phys.* **2016**, *144*, 214110.

(50) Chai, J.-D.; Head-Gordon, M. Systematic optimization of long-range corrected hybrid density functionals. *J. Chem. Phys.* **2008**, *128*, 084106.

(51) Mardirossian, N.; Head-Gordon, M. ωB97X-V: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy. *Phys. Chem. Chem. Phys.* **2014**, *16*, 9904–9924.

(52) Brémond, E.; Adamo, C. Seeking for parameter-free double-hybrid functionals: The PBE0-DH model. *J. Chem. Phys.* **2011**, *135*, 024106.

(53) Grimme, S. Semiempirical hybrid density functional with perturbative second-order correlation. *J. Chem. Phys.* **2006**, *124*, 034108.

(54) Brémond, E.; Sancho-García, J. C.; Pérez-Jiménez, A. J.; Adamo, C. Communication: Double-hybrid functionals from adiabatic-connection: The QIDH model. *J. Chem. Phys.* **2014**, *141*, 031101.

(55) Tarnopolsky, A.; Karton, A.; Sertchook, R.; Vuzman, D.; Martin, J. M. L. Double-Hybrid Functionals for Thermochemical Kinetics. *J. Phys. Chem. A* **2008**, *112*, 3–8.

(56) Brémond, E.; Pérez-Jiménez, A. J.; Sancho-García, J. C.; Adamo, C. Range-separated hybrid density functionals made simple. *J. Chem. Phys.* **2019**, *150*, 201102.

(57) Casanova-Páez, M.; Dardis, M. B.; Goerigk, L. ωB2PLYP and ωB2GPPLYP: The First Two Double-Hybrid Density Functionals with Long-Range Correction Optimized for Excitation Energies. *J. Chem. Theory Comput.* **2019**, *15*, 4735–4744.

(58) Brémond, E.; Savarese, M.; Pérez-Jiménez, A. J.; Sancho-García, J. C.; Adamo, C. Range-Separated Double-Hybrid Functional from Nonempirical Constraints. *J. Chem. Theory Comput.* **2018**, *14*, 4052–4062.

(59) Bayes, T. LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Phil. Trans. R. Soc.* **1763**, *53*, 370–418.

(60) Frazier, P. I. A Tutorial on Bayesian Optimization. 2018.

(61) Wu, Y.; Walsh, A.; Ganose, A. M. Race to the bottom: Bayesian optimisation for chemical problems. *Digit. Discov.* **2024**, *3*, 1086–1100.

(62) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(63) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA, 2016; pp 785–794.

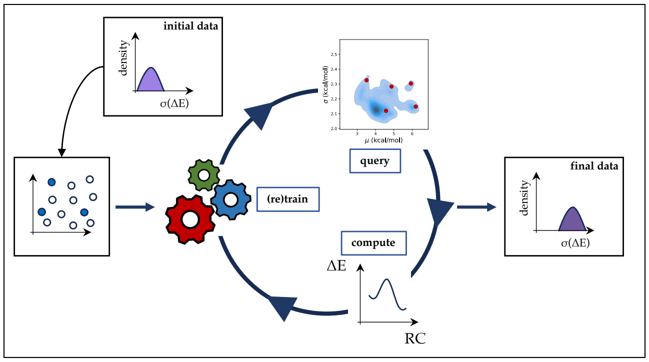
(64) Lan, Y.; Zou, L.; Cao, Y.; Houk, K. N. Computational Methods To Calculate Accurate Activation and Reaction Energies of 1,3-Dipolar Cycloadditions of 24 1,3-Dipoles. *J. Phys. Chem. A* **2011**, *115*, 13906–13920.

(65) Tang, S.-Y.; Shi, J.; Guo, Q.-X. Accurate prediction of rate constants of Diels–Alder reactions and application to design of Diels–Alder ligation. *Org. Biomol. Chem.* **2012**, *10*, 2673–2682.

(66) Casetti, N.; Alfonso-Ramos, J. E.; Coley, C. W.; Stuyver, T. Combining Molecular Quantum Mechanical Modeling and Machine Learning for Accelerated Reaction Screening and Discovery. *Chem. - Eur. J.* **2023**, *29*, e202301957.

(67) `https://github.com/chimie-paristech-CTM/ML_DFT_benchmarking`.

TOC Graphic


initial data
density
σ(ΔE)
2.5
2.4
2.3
2.2
2.1
2.0
σ (kcal/mol)
3
4
5
6
μ (kcal/mol)
query
(re)train
compute
ΔE
RC
final data
density
σ(ΔE)